ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202211.00424

Comparative Evaluation and Comprehensive Analysis of Machine Learning Models for Regression Problems: Postprint

Authors: Boran, Sekeroglu, Yoney, Kirsal Ever, Kamil, Dimililer, Fadi, Al-Turjman, Boran, Sekeroglu

Date: 2022-11-28T00:00:00+00:00

Abstract

Artificial intelligence and machine learning applications are of significant importance almost in every field of human life to solve problems or support human experts. However, the determination of the machine learning model to achieve a superior result for a particular problem within the wide real-life application areas is still a challenging task for researchers. The success of a model could be affected by several factors such as dataset characteristics, training strategy and model responses. Therefore, a comprehensive analysis is required to determine model ability and the efficiency of the considered strategies. This study implemented ten benchmark machine learning models on seventeen varied datasets. Experiments are performed using four different training strategies 60:40, 70:30, and 80:20 hold-out and five-fold cross-validation techniques. We used three evaluation metrics to evaluate the experimental results: mean squared error, mean absolute error, and coefficient of determination (R2 score). The considered models are analyzed, and each model' s advantages, disadvantages, and data dependencies are indicated. As a result of performed excess number of experiments, the deep Long-Short Term Memory (LSTM) neural network outperformed other considered models, namely, decision tree, linear regression, support vector regression with a linear and radial basis function kernels, random forest, gradient boosting, extreme gradient boosting, shallow neural network, and deep neural network. It has also been shown that cross-validation has a tremendous impact on the results of the experiments and should be considered for the model evaluation in regression studies where data mining or selection is not performed.

ChinaRxiv 中X

Full Text

Preamble

RESEARCH PAPER

Comparative Evaluation and Comprehensive Analysis of Machine Learning Models for Regression Problems

Boran Sekeroglu[1,5]†, Yoney Kirsal Ever[2,5], Kamil Dimililer[3,5], Fadi Al-Turjman[4,5]

[1]Information Systems Engineering Department, Near East University, Nicosia, Cyprus, Mersin 10, Turkey
[2]Software Engineering Department, Near East University, Nicosia, Cyprus, Mersin 10, Turkey
[3]Electrical and Electronic Engineering Department, Near East University, Nicosia, Cyprus, Mersin 10, Turkey
[4]Artificial Intelligence Engineering Department, Near East University, Nicosia, Cyprus, Mersin 10, Turkey
[5]Research Centre for AI and IoT, Near East University, Nicosia, Cyprus, Mersin 10, Turkey

Keywords: Machine learning; Regression; Comparative evaluation; Analysis; Validation

Citation: Sekeroglu, B., et al.: Comparative Evaluation and Comprehensive Analysis of Machine Learning Models for Regression Problems. Data Intelligence 4(3), 620-652 (2022). doi: 10.1162/dint_a_{00155}
Received: Jan. 10, 2022; **Revised:** Mar. 15, 2022; **Accepted:** Apr. 11, 2022

ABSTRACT

Artificial intelligence and machine learning applications have become critically important across nearly every domain of human life, either solving problems directly or supporting human experts. However, selecting the most appropriate machine learning model to achieve superior performance for a specific problem remains a significant challenge for researchers. A model' s success can be influenced by multiple factors, including dataset characteristics, training strategy, and model responsiveness. Consequently, comprehensive analysis is essential to determine model capability and the effectiveness of employed strategies. This study implemented ten benchmark machine learning models across seventeen diverse datasets using four training strategies: 60:40, 70:30, and 80:20 hold-out splits, along with five-fold cross-validation.

We evaluated experimental results using three metrics: mean squared error, mean absolute error, and coefficient of determination (R^2 score). Each model was analyzed to identify its advantages, disadvantages, and data dependencies. Based on extensive experimentation, the deep Long-Short Term Mem-

ory (LSTM) neural network outperformed all other considered models, including decision tree, linear regression, support vector regression with linear and radial basis function kernels, random forest, gradient boosting, extreme gradient boosting, shallow neural network, and deep neural network. Additionally, cross-validation demonstrated tremendous impact on experimental outcomes and should be considered essential for model evaluation in regression studies where data mining or selection is not performed.

† **Corresponding Author:** Boran Sekeroglu (E-mail: boran.sekeroglu@neu.edu.tr; ORCID: 0000-0001-7284-1173)

1. INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) models excel at establishing relationships between variables (attributes) and observations (instances) across various tasks, including classification and regression. Unlike classification, where samples are assigned discrete labels, regression aims to fit a line or plane through all samples with minimal error. In several studies, researchers have treated real-valued outputs as bucketed categories, converting regression problems into classification tasks when possible [**?**]. This has led to more ML implementations for classification than for regression, which seeks to predict continuous, infinite outputs. Nevertheless, regression applications have significantly impacted multidisciplinary fields, including healthcare [**?**, **?**], education [**?**], price prediction [**?**], sports [**?**], and finance [**?**].

The primary challenge in ML applications is identifying the model most suitable for a particular dataset. This is crucial because determining a universally optimal ML model for all applications is nearly impossible due to varying dataset characteristics and model capabilities [**?**]. Dataset properties further complicate research in both domains. Based on attribute characteristics, datasets may be structured or unstructured, numeric or categorical, or combined. Most importantly, relationships between attributes and outputs can be linear or nonlinear, with high or low correlation, causing ML models to produce different results across datasets. Therefore, analyzing model success requires testing across datasets with diverse conditions.

Beyond dataset characteristics, training strategies (validation techniques) and evaluation methods vary considerably across studies. The hold-out method, which splits data into training and testing sets at different ratios (60:40, 70:30, 75:25, 80:20, etc.), is common in AI and ML implementations [**?**, **?**, **?**]. Another widely used approach is k-fold cross-validation, employed for hyperparameter tuning and final evaluation [**?**]. The main drawback of hold-out is that samples are used exclusively for either training or testing, making model performance dependent on the specific split. In contrast, k-fold cross-validation provides more accurate results [**?**] by partitioning data into k equal folds and training models k times, ensuring all data contributes to both training and testing. However,

using all samples can affect results positively or negatively, slightly or significantly. For models trained with random data selection without data mining, the amount of training data and validation method significantly impact outcomes. Recent regression studies vary in model selection, evaluation, and training implementation, though comparing multiple models across several metrics is common [**?**, **?**, **?**, **?**, **?**].

Even in recent research, no standard exists for hold-out ratios or cross-validation implementation. Choices depend on dataset size, researcher preference, and model response. However, in ML, even small changes in training data can significantly affect results, particularly for datasets with different characteristics or large-scale data. Therefore, investigating the effects of both cross-validation and hold-out ratios across varied datasets is crucial [**?**].

While most ML studies include comparative analysis, few conduct direct comparisons. Huang et al. [**?**] compared three ML models (backpropagation neural network, support vector regression, and extreme learning machine) for regression using four datasets, evaluating with mean squared error (MSE), mean absolute percentage error (MAPE), and R^2 score using 20-fold cross-validation. They concluded that integrated models outperformed single models. Bratsas et al. [**?**] compared multilayer perceptron, linear regression, random forest, and support vector regression for predicting traffic status in Thessaloniki, Greece, using three scenarios from a single dataset and root mean square error (RMSE) for evaluation. They found neural networks and support vector regression outperformed random forest and linear regression. Recent comparative research demonstrates that analyzing model capabilities and training strategy effects requires multiple diverse datasets, models, and validation techniques.

Automated Machine Learning (AutoML) has recently emerged to identify superior models among various ML approaches and achieve optimal results through ensemble methods. While AutoML offers implementation advantages, its computational costs and tendency to cause system crashes, even on relatively small datasets, remain major disadvantages.

This study compares ML models for regression tasks using scenarios not previously examined together. We consider numerous multi-character datasets—including time-series, multivariate, high-instance, and high-attribute data—across varied validation strategies to analyze model responses to different training data volumes and the effects of hold-out versus cross-validation on regression tasks. Our primary goal is to establish starting points for future regression studies to minimize model and validation strategy selection efforts.

To this end, we selected ten benchmark ML models based on their frequency of use and foundational importance. Linear Regression remains one of the most frequently used statistical models for regression, particularly for linearly related data. Decision Tree (DT) is another common model that forms the basis for tree-ensemble models such as Random Forest (RF), Gradient Boosting (GradBoost), and Extreme Gradient Boosting (XGBoost). These ensemble methods

minimize DT error through bagging or boosting and are increasingly popular for regression. Support Vector Regression (SVR) extends the success of support vector machines to regression problems, though kernel selection presents implementation challenges; we consider both Radial Basis Function (RBF) and linear kernels. Neural networks are essential tools for nonlinear data, so we implement both shallow and deep versions (NN and DNN), plus a specialized recurrent architecture: deep Long-Short Term Memory (deep LSTM).

We conducted 680 experiments across 17 datasets for comprehensive evaluation. Results were analyzed using three common regression metrics: MSE, MAE, and R^2 scores. We examined hold-out method effects at different ratios, analyzed performance changes with varying training data, determined model data dependencies, compared hold-out results to five-fold cross-validation, performed fold-level analysis with statistical descriptions, conducted model-based evaluations, and presented advantages and disadvantages of each approach. Finally, we offer recommendations for model and validation strategy selection.

2.1 Datasets

We selected 17 regression datasets from diverse real-world domains—environmental sciences, social sciences, civil engineering, finance, sales, and energy consumption—to enable generalizable comparisons across application fields. Datasets varied in attribute and instance counts to assess model capabilities across different data scales and types, including time-series and multivariate data.

Specifically, we used: Air Quality [**?**], Wine Quality [**?**], Combined Cycle Power Plant (CCPP) [**?**], Behavior of urban traffic in Sao Paulo, Brazil (SPB) [**?**], Real Estate (RE) Valuation [**?**], Concrete Compressive Strength (CON) [**?**], Daily Demand Forecasting Orders (DDFO) [**?**], two Student Performance (SP) datasets [**?**], and three Power Consumption of Tetouan city (TCPC) datasets [**?**] (one per zone). summarizes instance and attribute counts.

Numerical representations obscure data relationships and characteristics. [Figure 1: see original paper] presents correlation analyses. The highest attribute correlations appear in AQ and DDFO datasets, while the lowest occur in WQR, WQW, RE, and STP datasets. For the STM dataset, we removed two highly correlated attributes to create a more challenging STP dataset.

2.2 Brief Review of Machine Learning Algorithms

2.2.1 Artificial Neural Networks

Backpropagation is the most widely used neural network for optimization, regression, and classification. Interconnection weights update based on the difference between actual and expected outputs, with gradient descent calculating weight

changes. It remains a standard algorithm for comparative studies [**?**]. We used a shallow version with one hidden layer and a deep version with four hidden layers.

2.2.2 Linear Regression

Linear Regression is a statistical method that fits the best regression line through data points. It is frequently and successfully applied to datasets with linear attribute correlations [**?**].

2.2.3 Support Vector Regression

Support Vector Regression (SVR) was developed to produce real-valued outputs for regression problems [**?**]. It minimizes error while maximizing hyperplane margin for effective data separation [**?**, **?**]. Different kernel functions project data into higher dimensions; we compared Linear and Radial-Basis Function kernels.

2.2.4 Long-Short Term Memory Neural Network

LSTM is an effective recurrent network variant for classification and regression [**?**]. Its architecture comprises four major components: cell, input gate, output gate, and forget gate. LSTM uses gradients to update weights while remembering previous errors, improving error minimization across iterations [**?**].

2.2.5 Decision Tree

Decision Trees are tree-structured algorithms with root, decision, and leaf nodes that employ a divide-and-conquer strategy, offering both advantages and disadvantages [**?**]. Simplicity and speed are primary benefits, while determining initial root nodes and attribute sequences represents the main drawback.

2.2.6 Random Forests

Random forests are tree-based ensemble methods applicable to classification and regression [**?**, **?**]. They construct multiple decision trees during training and optimize the mean prediction across individual trees.

2.2.7 Gradient Boosting Algorithm

Gradient Boosting is another tree-based ensemble algorithm [**?**]. It optimizes outputs by minimizing loss from weak learners (decision trees). Loss is calculated, and new or modified trees are added to reduce total loss via gradient descent. Output is modified after each tree addition, with stopping criteria such as loss plateau or fixed tree count determining the final model.

2.2.8 Extreme Gradient Boosting

Similar to Gradient Boosting, Extreme Gradient Boosting [**?**] is an ensemble tree method that boosts weak learners using gradient descent. However, XGBoost includes enhancements to minimize resource usage and improve results. Regularization models (e.g., LASSO) prevent overfitting, and built-in cross-validation determines optimal iteration counts in a single run.

2.3 Evaluation and Comparison Criteria

We used three common regression metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and coefficient of determination (R^2 score).

MSE squares errors before averaging, giving higher weight to significant outliers. This helps researchers observe errors in high-value datasets, though error frequency substantially affects MSE results, with repeated errors causing MSE to increase. However, minimum error does not always indicate superior predictions, as significant errors may inflate MSE and cause overestimation of model errors. Conversely, small errors may underestimate total error. Therefore, multiple evaluation criteria must be considered during model assessment.

MAE, the mean of absolute errors, focuses on error magnitude between predicted and actual outputs without considering direction, yielding more stable results. The R^2 score, strongly related to MSE, measures correlation between predicted and observed values, providing scaled evaluation results for robust model comparison.

2.4 The Design of Relevant Experiments

Experiments employed four training configurations: three hold-out ratios (60:40, 70:30, 80:20) and five-fold cross-validation. Models were trained on each hold-out ratio, with scores obtained from separate test sets. Five-fold cross-validation provided more accurate and robust evaluation by averaging results across folds.

These results enabled analysis of how hold-out and cross-validation strategies affect scores, using both overall hold-out ratios and individual fold results. Neural network architectures (NN, DNN, deep LSTM) were fixed, though parameters were tuned per dataset performance.

For hold-out experiments, fixed training and testing data eliminated performance variation from data changes. In cross-validation, each fold was fixed to ensure all models trained on identical data splits.

DNN implementation used four hidden layers with 500 neurons each, Sigmoid activation, Adam optimizer, and MSE loss. The shallow NN used one hidden layer with 500 neurons and identical other parameters. Deep LSTM used four

layers, with maximum iterations determined by highest achieved scores. DNN, NN, and LSTM experiments were repeated with different iterations per dataset to optimize results.

Grid search optimized parameters for Support Vector Regression, Random Forest, GradBoost, and XGBoost. Best parameters were used for each cross-validation fold and all hold-out ratios. MSE was used to build decision tree regressor structures.

---

3. RESULTS AND COMPARISONS

This section summarizes experimental results and compares models across training ratios with quantitative data. All models produced fluctuating results across learning rates and datasets. Detailed comparisons follow, with S1 Table, S2 Table, and S3 Table presenting MSE, MAE, and R^2 scores for all experiments. [Figure 2: see original paper] visualizes R^2 scores per dataset for each hold-out ratio and cross-validation. Bold values indicate superior results.

3.1 Comparisons for Hold-Out Ratios and Cross-Validation

Analyzing hold-out ratio effects requires two-stage examination: first, assessing performance changes with training data volume; second, determining how different ratios and cross-validation impact specific models, revealing data dependency and sensitivity.

R^2 scores were used for analysis due to their scaled nature, enabling more effective evaluation. We examined both highest R^2 scores and statistical descriptions (mean, median, quartiles, standard deviation) per dataset and model.

NN, DT, and RF produced fluctuating but similar results across hold-out ratios, with lowest and highest R^2 scores typically at 70:30 and 80:20, respectively. GradBoost and XGBoost also fluctuated, achieving lowest scores at 70:30 but highest at 60:40.

LSTM, SVRL, SVRBF, and LR showed negative linear relationships between training data volume and performance, with highest R^2 scores at 60:40 and decreasing results as training data increased, reaching minima at 80:20.

DNN was most positively affected by training data increases. It achieved no highest R^2 scores at 60:40 but improved at 70:30 and 80:20, peaking at 80:20.

Cross-validation significantly changed results for NN, DNN, LR, DT, GradBoost, and XGBoost, enabling these models to achieve superior results. However, SVRL, SVRBF, and RF showed no positive cross-validation impact.

Statistical descriptions (mean, median, min/max, standard deviation, quartiles) reveal model sensitivity to training data volume and adaptation. Deep LSTM achieved superior results across all statistical measures and scenarios. SVRL

and LR produced highest standard deviations and lowest mean/median results, representing the worst performance.

[Figure 3: see original paper] shows model-based result distributions, while [Figure 4: see original paper] presents validation strategy comparisons per model.

NN and DNN showed minimal standard deviation variation across ratios (NN: 0.0006–0.0612; DNN: 0.0003–0.0641), representing the lowest maximum standard deviations. LR followed with 0.0797 maximum standard deviation and lowest average standard deviation (0.0215), despite rarely achieving superior results. SVRBF, SVRL, and DT showed similar maximum standard deviations, with averages of 0.0154, 0.0173, and 0.0275, respectively—SVR models having the lowest average standard deviation.

Tree ensemble models (RF, GradBoost, XGBoost) showed greater fluctuation, with maximum standard deviations of 0.1384, 0.1569, and 0.1461, and averages of 0.0305, 0.0322, and 0.0331. Deep LSTM exhibited the largest standard deviation changes, with maximum and average values of 0.3467 and 0.0629—the highest among all models. summarizes minimum, maximum, and average standard deviations.

3.2 Fold Comparisons of Five-Fold Cross-Validation Experiments

Five-fold cross-validation analyzed how changing training data affects model learning and performance, demonstrating sensitivity compared to fixed hold-out ratios. We used average (difference between highest and lowest fold R^2 scores) to indicate general variation.

AQ and CCPP datasets were excluded from this section because all models produced R^2 scores above 0.999 with less than 0.001 average variation between folds.

NN showed highest variation (= 0.48) in SPB dataset (fold range: 0.81 to 0.32). It produced stable results in WQW, CON, DDFO, and STM (= 0.07-0.10) but more fluctuation in WQR, RE, and STP (= 0.17, 0.21, 0.20). DNN produced similar but more stable results than NN in WQR, CON, STM, and STP, with maximum variation again in SPB (= 0.47) and average = 0.16.

LR achieved perfect consistency in DDFO (= 0) but failed to produce results in one SPB fold (= 0.74). Its most stable performance outside DDFO was WQW (= 0.15), with overall average = 0.26.

SVRL and SVRBF produced similar values across most datasets, with maximum differences in CON (SVRL = 0.51, SVRBF = 0.44) and STP (SVRL = 0.26, SVRBF = 0.30). Both showed high variability in SPB (= 0.73 each), with average = 0.29.

Deep LSTM, despite superior overall results, produced the most fluctuating, data-dependent results in DDFO, STM, and STP. While showing minimal variation in most datasets and relatively stable SPB performance, extreme predic-

tions in STM and STP folds drove to 0.99. Its average = 0.36 marks it as the most sensitive model.

DT failed to produce results in any STP fold (= 0, excluded from average) and missed single folds in WQR and WQW (= 0.13 and 0.21). It achieved relatively stable SPB results (= 0.23) despite not being superior, but showed highest variation in RE (= 0.37, range: 0.69–0.32). Average = 0.22.

RF, though more stable than DT, missed folds in CON and SPB, causing high variation (= 0.83 each) and average = 0.33. GradBoost produced similar but more stable results in CON and SPB, achieving average = 0.17 as one of the most stable models.

XGBoost achieved the lowest average (0.15), with minimum variation in WQW and STP (= 0.01 and 0.11) and maximum in RE and SPB (= 0.30 and 0.29). It generally produced the most stable results despite rarely achieving superiority. [Figure 5: see original paper] plots fold-wise R^2 score variations, and presents values per model and dataset.

---

4. DISCUSSIONS

Multiple aspects warrant discussion regarding the obtained results. We separately analyzed model performance by R^2 scores and error minimization across dataset types, examined hold-out training ratio effects, and assessed cross-validation impact on model learning to derive general conclusions. Data dependency was analyzed through combined results and fold analysis, providing insights into model consistency and stability.

4.1 Effect of Training Ratios on Model Performances

Comparing varied hold-out ratios without cross-validation yields inconsistent analysis due to model responses to newly added training data, even with identical samples. Fluctuations occurred regardless of training ratios, though ratio-based changes were not always significant.

SVRBF and SVRL were least affected by training data volume, indicating that data projection reduces sensitivity to sample size. SVRBF achieved the lowest average standard deviation, followed by SVRL. However, increasing training data negatively impacted both models, reaffirming that SVR models can perform better with less training data.

Neural-based models (NN, DNN) also minimized variation through effective convergence enabled by hidden layers and neurons. Though less interpretable, they achieved stable results. However, DNN with more layers and neurons showed increased success rates with more training data, demonstrating that deeper architectures require larger datasets. Fewer hidden layers and neurons in NN produced more inconsistent results, complicating optimal ratio determination.

DT's node sequence and leaf determination critically affect its success, producing fluctuating results across training ratios. However, RF reduced DT' s need for more training data, while GradBoost and XGBoost eliminated it entirely.

RF, GradBoost, and XGBoost showed higher standard deviations than other models, indicating greater sensitivity to training data volume. Yet their processes minimized DT errors and reduced dependence on data quantity, enabling GradBoost and XGBoost to achieve higher results with less training data.

Deep LSTM was most sensitive to training data volume, producing the most fluctuating results in terms of both R^2 scores and standard deviation. While fixed LSTM layers may contribute to this, the forgetting and remembering of input sequence states significantly affected both success and variability. Despite using many LSTM layers, deep LSTM achieved highest results with the lowest training ratio (60:40), with increased ratios not improving performance.

Across all models, the 70:30 ratio produced minimum success, while 80:20 yielded higher results, but 60:40 achieved the highest overall scores. If cross-validation is not considered, using 60:40 or 80:20 ratios based on model characteristics can reduce experimental costs while achieving superior results.

4.2 Effect of Cross-Validation and Data Dependency of the Models

Cross-validation is frequently used for hyperparameter tuning and final evaluation. Our results demonstrate its vital role in determining model generalization ability. Five-fold cross-validation effectively trained models five times at 80:20 ratios using different data partitions, enabling comprehensive analysis.

Average fold results showed NN, DNN, LR, DT, GradBoost, and XGBoost achieved their highest R^2 scores with cross-validation (S3 Table), indicating that cross-validation provides more consistent, reliable, and successful outcomes for these models. However, SVRBF, SVRL, and LSTM showed no remarkable improvement over other ratios.

Examining fold-wise R^2 score differences ( ) revealed more complex relationships (TABLE:4). values indicated model responses to training data changes and data dependency. XGBoost was most successful, with minimum data dependency (min $= 0.01$, max $= 0.31$, average $= 0.113$). By minimizing error while adding trees, XGBoost reduced dependence on new data, producing stable inter-fold results.

GradBoost, DNN, and NN followed XGBoost. GradBoost and DNN had equal average (0.17). Similar to XGBoost, GradBoost' s ensemble creation via gradient descent and weak trees yielded low data dependency. DNN' s many hidden layers and neurons produced stable results despite not being superior, enhancing learning from changing cross-validation data. NN' s single hidden layer produced more fluctuation than DNN, but minimal R^2 score changes between folds showed that increasing hidden layers improves stability without significantly altering overall results.

DT and LR produced similar values. Both excel on linearly related datasets, increasing data importance in training and testing. However, LR was more stable on highly correlated data, while DT showed greater fold-wise R^2 differences. Single tree creation and node sequences were critical factors.

SVRBF and SVRL produced reasonably similar results (= 0.276 and 0.251). High fold variation stemmed from unpredictable data projection—while one fold' s data might project well for regression, another might not, creating variable results.

RF and LSTM produced unexpected results. As an ensemble, RF should have been stable, but tree creation effects made it highly data-dependent despite successful results. LSTM, though generally stable, produced extreme results in STP and STM folds, yielding the highest average (0.36). While LSTM appears highly data-dependent, dataset characteristic analysis follows.

4.3 Effect of Datasets on Model Performances

Dataset-based analysis evaluated model responses under varied conditions. While large instance counts with many attributes facilitate learning (e.g., AQ dataset), instance relevancy and information quality are vital (e.g., WQW dataset). Neural networks, expected to excel on nonlinear problems with weak attribute relationships, only slightly outperformed or underperformed other models on these datasets.

On STM and STP datasets, removing two highly correlated attributes severely reduced NN and DNN success, similar to other models. While NN and DNN produced high results on highly correlated DDFO data, they did not match LR' s prediction rate. However, outperforming LR on the highly correlated AQ dataset demonstrated that minimized training instances significantly and negatively affected NN and DNN performance.

SVRBF and SVRL achieved results close to NN and DNN but excelled on highly correlated datasets with minimal attributes and instances (e.g., DDFO). On other datasets with more attributes and instances, even with high correlation, SVR models underperformed NN and DNN, reaffirming that limited, informative attributes and instances enhance SVR success.

DT produced fluctuating, generally unsuccessful results depending on dataset characteristics, failing entirely on the high-attribute, low-instance, low-correlation STP dataset. However, DT outperformed RF, XGBoost, and GradBoost on the highly correlated DDFO dataset, showing that success in identifying starting nodes and attribute sequences is vital.

Tree-based ensemble models achieved superior results on high-instance datasets. RF, GradBoost, and XGBoost generally outperformed others (except LSTM) on AQ, CCPP, and WQW datasets, as many instances enabled systematic information connections through tree creation and addition.

LSTM outperformed all models regardless of attribute correlation when trained with sufficient instances. However, LSTM was most affected by instance and attribute counts, producing highly fluctuating results on STM, STP, and DDFO datasets. Training deep LSTM with limited data makes achieving high prediction rates challenging.

4.4 Discussions on the General Results

Deep LSTM achieved the highest R^2 scores and superior error minimization in 57 of 68 experiments, demonstrating exceptional success. GradBoost followed, producing twice as many superior results and three times as many second-highest results. XGBoost achieved one superior and ten second-highest results, showing stable performance. LR produced superior results once, as did DT and RF. NN, DNN, and SVR models never achieved optimal results.

Although DNN' s additional hidden layers and neurons improved stability over NN, it never achieved superior or even second-superior scores, limiting neural-based model success in this study. However, fixed architectures and parameters may explain this limitation. NN' s reduced hidden layers and neurons negatively affected stability without drastically reducing success compared to DNN, decreasing reliability.

SVR models produced results similar to but slightly lower than neural-based models. While they achieved higher, more stable results than NN and DNN on low-instance, low-attribute datasets, they generally lagged behind LSTM and tree-ensemble models.

DT underperformed compared to tree-based ensembles. Optimizing multiple trees inevitably yields superior results over single trees, which motivated ensemble development. Though easily implementable and responsive, DT' s limitation —attribute sequence determination—was overshadowed by RF, GradBoost, and XGBoost success.

LR, the most basic regression model, reaffirmed its success on linear problems and showed capability to correlate correctly selected data with nonlinear test data. However, its limited ability to establish nonlinear correlations across all datasets requires further experimentation. LR remains essential for linearly related attributes.

RF slightly underperformed GradBoost and XGBoost among tree-based models. While RF optimizes tree regression results through bagging, GradBoost and XGBoost consider weak sample losses through boosting, making them less sensitive to overfitting and more advantageous.

GradBoost and XGBoost proved to be models that should be prioritized for regression problems, offering stability, low data dependency, and strong results. Parameter optimization could further enhance their performance.

Deep LSTM achieved the highest results in nearly all experiments, even on

datasets where other models failed, establishing it as a top regression model. However, its significant disadvantages include high data dependency, unstable responses to changing training data, and inconsistent results under some conditions. Between folds, LSTM could predict all data perfectly in one fold and fail completely in another. Forgetting states' dependence on previous data and effects on subsequent sequences significantly impacted consistency. Fixed LSTM structure may cause fluctuations, requiring further experiments for generalization.

Given that all experiments used randomly selected then fixed data, fluctuations and instability in many models were acceptable. LSTM' s major limitation was failing to complete learning on low-instance, high-attribute datasets (STM and STP). With appropriate and sufficient data selection, deep LSTM should continue achieving high-level regression results.

4.5 Outcomes and Recommendations for Further Studies

- LSTM, XGBoost, and GradBoost were superior overall and should be widely used for regression problems.
- RF, SVR models, and LSTM were most data-dependent, making cross-validation essential for evaluating their prediction abilities.
- Optimal hold-out ratios vary by data and model, but 60:40 achieved the highest results overall.
- XGBoost showed least dependence on training data volume and changes, followed by GradBoost. Their results are robust to hold-out or cross-validation choice.
- NN and DNN were minimally affected by data changes, but cross-validation remains valuable for parameter and structure determination.
- Deeper neural networks outperform shallow ones as data volume (instances and/or attributes) increases.
- LSTM achieved superior results across varied dataset types and sizes, but structural determination can cause inconsistency.
- XGBoost and GradBoost, while not matching LSTM' s peak performance, produced stable, data-independent results at low computational cost.
- Cross-validation affected all models' R^2 scores by 0.01–0.20 on average, proving essential for consistent, comparable results even for data-independent models.
- Unselected data feeding to SVR models prevented superior performance; data minimization might enhance SVR ability.
- LR remains superior for linear relationships but struggles with complex datasets.
- AutoML approaches should consider model characteristics from this study to reduce computational costs and offer more effective implementations.

4.6 Limitations of the Study

This study has limitations. Using only two SVR kernels, implementing fixed structures for deep LSTM, NN, and DNN, and optimizing these aspects could improve results. Additionally, exploring various k values in k-fold cross-validation would provide further insights into model behavior.

---

5. CONCLUSION

Comparing machine learning models is challenging due to numerous models, diverse datasets, and varied training strategies. This study performed 680 experiments across 15 datasets, training ten benchmark models with three hold-out ratios and five-fold cross-validation.

Results show each model has unique weaknesses and strengths for regression implementation. Despite significant data dependency and sensitivity to training data changes, deep LSTM significantly outperformed other models in nearly all experiments. Linear Regression remains essential for highly correlated data. Tree-based ensembles, particularly Gradient Boosting and XGBoost, provide reliable, consistent results with low sensitivity to training data changes. Neural-based models produced stable but non-superior results, requiring more data for deeper architectures. SVR models achieved superior results only with minimal attributes and instances, but their data dependency complicates implementation.

Different hold-out ratios did not significantly affect performance, with 60:40 being most beneficial. However, cross-validation training considerably impacts prediction ability and should be used instead of fixed, randomized training ratios for robust analysis.

DATA AVAILABILITY

The data supporting this study are openly available in the UCI Machine Learning Repository at https://archive.ics.uci.edu/ \cite{14-22}.

AUTHOR CONTRIBUTIONS

Boran Sekeroglu (boran.sekeroglu@neu.edu.tr): Conceptualization, Methodology, Software, Writing–original draft preparation, Writing–review and editing.
Yoney Kirsal Ever (yoneykirsal.ever@neu.edu.tr): Conceptualization, Project administration, Methodology, Writing–original draft preparation, Writing–review and editing.
Kamil Dimililer (kamil.dimililer@neu.edu.tr): Conceptualization, Methodology, Software, Writing–original draft preparation, Writing–review and editing.
Fadi Al-Turjman (fadi.alturjman@neu.edu.tr): Conceptualization, Methodology, Writing–review and editing.

REFERENCES

[1] Ever, Y.K., Dimililer, K., Sekeroglu, B.: Comparison of Machine Learning Techniques for Prediction Problems. In Advances in Intelligent Systems and Computing 927, 713–723 (2019)
[2] Sekeroglu, B., Tuncal, K.: Prediction of cancer incidence rates for the European continent using machine learning models. Health Informatics Journal 27(1), 1460458220983878 (2021)
[3] Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., Pinheiro, P.R.: CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection. IEEE Access 8, 91916–91923 (2020)
[4] Mesaric, J., Sebalj, D.: Decision trees for predicting the academic success of students. Croatian Operational Research Review 7, 367–388 (2016)
[5] Utomo, D., Pujiono, S.M., et al.: Stock price prediction using back propagation neural network based on gradient descent with momentum and adaptive learning rate. Journal of Internet Banking and Commerce 22, 1–16 (2017)
[6] Oytun, M., Tinazci, C., Sekeroglu, B., Acikada, C., Yavuz, H.U.: Performance prediction and evaluation in female handball players using machine learning models. IEEE Access 8, 116321–116335 (2020)
[7] Taboga, M.: Cross-country differences in the size of venture capital financing rounds: a machine learning approach. Empirical Economics 5 (2021)
[8] Dougherty, G.: Pattern Recognition and Classification. Springer (2013)
[9] Pekel, E.: Estimation of soil moisture using decision tree regression. Theoretical and Applied Climatology 139 (2020)
[10] Pandey, S., Kumar, V., Kumar, P.: Application and analysis of machine learning algorithms for design of concrete mix with plasticizer and without plasticizer. Journal of Soft Computing in Civil Engineering 5(1), 19–37 (2021)
[11] Kaveh, A., Eslamlou, A.D., Javadi, S.M., Malek, N.G.: Machine learning regression approaches for predicting the ultimate buckling load of variable-stiffness composite cylinders. Acta Mechanica 1–11 (2021)
[12] Huang J.C., Ko K.M., Shu M.H., Hsu B.M.: Application and comparison of several machine learning algorithms and their integration models in regression problems. Neural Computing and Applications 32, 5461–5469 (2020)
[13] Bratsas, C., Koupidis, K., Salanova, J.M., Giannakopoulos, K., Kaloudis, A., Aifadopoulou, G.: A comparison of machine learning methods for the prediction of traffic speed in Urban Places. Sustainability 12(1) (2020)
[14] DeVito, S., Massera, E., Francia, G. Di, Piga, M., Martinotto, L:. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario: Sensors and Actuators B. Chemical 129(2), 750–757 (2008)
[15] Zhong, P., Fukushima, M.: Regularized non-smooth newton method for multi-class support vector machines. Methods and Software 22(1), 225–236 (2007)
[16] Tufekci, P.: Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. International Journal of Electrical Power and Energy Systems 60, 126–140 (2014)

[17] Ferreira, R.P., Affonso, C., Sassi, R.J.: Combination of artificial intelligence techniques for prediction the behavior of urban vehicular traffic in the city of Sao Paulo. In 10th Brazilian Congress on Computational Intelligence (CBIC), pp. 1–7 (2011)
[18] Yeh, I.C., Hsu, T.K.: Building real estate valuation models with comparative approach through case-based reasoning. Applied Soft Computing 65, 260–271 (2018)
[19] Yeh, I.-C.: Modeling of strength of high-performance concrete using artificial neural networks. Cement and Concrete Research 28(12), 1797–1808 (1998)
[20] Ferreira, R.P., Martiniano, A., Ferreira, A., Ferreira, A., Sassi, R.J.: Study on daily demand forecasting orders using artificial neural network. IEEE Latin America Transactions 14(3), 1519–1525 (2016)
[21] Cortez, P., Silva, A.: Using data mining to predict secondary school student performance. In EUROSIS (2008)
[22] Salam, A.R., Hibaoui, A.E.: Comparison of machine learning algorithms for the power consumption prediction: case study of Tetouan city. In 2018 6th International Renewable and Sustainable Energy Conference (IRSEC), pp. 1–5, (2018)
[23] Amirjanov, A., Dimililer, K.: Image compression system with an optimisation of compression ratio. IET Image Processing 13(11), 1960–1969 (2019)
[24] Eyvazian, M., Noorossana, R., Amiri, A.S.A., et al.: Phase II monitoring of multivariate multiple linear regression profiles. Quality and Reliability Engineering International 27(3), 281–296 (2011)
[25] Smola, A.J., Scholkopf, B.: A tutorial on support vector regression. Statistics and Computing 14(3), 199–222 (2004)
[26] Henrique, B.M., Sobreiro, V.A., Kimura, H., et al.: Stock price prediction using support vector regression on daily and up to the minute prices. The Journal of Finance and Data Science 4(3), 183–201 (2018)
[27] Azeez, O.S, Pradhan, B., Shafri, H.Z.M., et al.: Vehicular CO emission prediction using support vector regression model and GIS. Sustainability 10(10) (2018)
[28] Ping, L., Jin, W., Sangaiah, A.K., Xie, Y., Yin, X., et al.: Analysis and prediction of water quality using LSTM deep neural networks in IoT environment. Sustainability 11, 2058 (2019)
[29] Wang, Y.: A new concept using LSTM Neural Networks for dynamic system identification. In American Control Conference (ACC), (2017)
[30] Yang, L., Wu, H., Jin, X., et al.: Study of cardiovascular disease prediction model based on random forest in eastern China. Scientific Reports 10, 5245 (2020)
[31] Pahlavan-Rad, M.R., Dahmardeh, K., Hadizadeh, M., et al.: Prediction of soil water infiltration using multiple linear regression and random forest in a dry flood plain, eastern Iran. CATENA 194, 104715 (2020)
[32] Friedman, J.: Greedy function approximation: A gradient boosting machine. Annals of Statistics 29, 1189–1232 (2001)
[33] Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. arXiv preprint arXiv:1603.02754 (2016)

SUPPORTING INFORMATION

S1 Table - All MSE results obtained in this study.
S2 Table - All MAE results obtained in this study.
S3 Table - All R^2 scores obtained in this study.

AUTHOR BIOGRAPHY

Boran Sekeroglu received his B.S., M.S., and Ph.D. degrees in computer engineering from Near East University, Nicosia, Cyprus, in 2001, 2004, and 2008, respectively. From 2009 to 2012, he was an Assistant Professor at the Computer Engineering Department, and currently, he is an Associate Professor at Near East University and serves as chairperson of the Information Systems Engineering Department. He has published over 60 peer-reviewed papers in journals and conferences related to machine learning, deep learning, and computer vision. He is a member of the Research Centre for AI and IoT, Applied Artificial Intelligence Research Center, and DESAM Research Institute. He reviews papers for journals, mainly on machine learning and deep learning. ORCID: 0000-0001-7284-1173

Yoney Kirsal Ever obtained her BSc. degree from the Department of Computer Engineering, Eastern Mediterranean University, Cyprus, in 2002, her MSc in Internet Computing from the University of Surrey, Guilford, Surrey, UK, in 2003, and her Ph.D. from the School of Engineering and Information Sciences at Middlesex University, London, UK. In 2012 she completed her Post-Graduate Certificate in Higher Education. Her research focuses on developing security strategies using Kerberos in wireless networks. Yoney has worked as a part-time lecturer during her BSc and Ph.D. studies and as a lecturer in the Computer and Communications Engineering Department at Middlesex University London. Currently, she is an Associate Professor and Chairperson in the Software Engineering Department at Near East University, Cyprus. She has published international conference papers with various awards, including IEEE best paper for promising research. Her research interests include network security, authentication protocols, and formal verification methods. Dr. Kirsal Ever has been a member of ACM since 2007 and a Member (M) of IEEE since 1998. She reviews papers for various journals, mainly on network security. ORCID: 0000-0002-8129-9846

Kamil Dimililer was born in Nicosia, Cyprus, in 1978. He received his B.Sc., M.Sc., and Ph.D. degrees in Electrical & Electronic Engineering from Near East University in 2002, 2004, and 2009–2014, respectively. He is an active researcher in the Applied Artificial Intelligence Research Centre (AAIRC) and a contributor to the International Research Center for AI and IoT at Near

East University. Currently, he is an Associate Professor and Chairperson of the Automotive Engineering Department. He has more than 100 publications in journals, conferences, and book chapters. He is an active reviewer for various journals. His research interests include Artificial Intelligence, Machine Learning, Pattern Recognition, Image Processing, Neural Networks, and Computer Vision. ORCID: 0000-0002-2751-0479

Prof. Dr. Fadi Al-Turjman received his Ph.D. in computer science from Queen's University, Canada, in 2011. He is the Associate Dean for Research and Founding Director of the International Research Center for AI and IoT at Near East University, Nicosia, Cyprus. Prof. Al-Turjman is Head of the Artificial Intelligence Engineering Department and a leading authority in smart/intelligent IoT systems, wireless and mobile networks'architectures, protocols, deployments, and performance evaluation in Artificial Intelligence of Things (AIoT). His publication history spans over 400 SCI/E publications, plus numerous keynotes and plenary talks at flagship venues. He has authored and edited more than 40 books on cognition, security, and wireless sensor networks in smart IoT environments, published by Taylor and Francis, Elsevier, IET, and Springer. He has received several recognitions and best paper awards at top international conferences, including the prestigious Best Research Paper Award from Elsevier Computer Communications Journal for 2015–2018, and the Top Researcher Award for 2018 at Antalya Bilim University, Turkey. Prof. Al-Turjman has led numerous international symposia and workshops in flagship communication society conferences. He currently serves as book series editor and lead guest/associate editor for several top-tier journals, including IEEE Communications Surveys and Tutorials (IF 23.9) and Elsevier Sustainable Cities and Society (IF 7.8), in addition to organizing international conferences and symposiums on cutting-edge AI and IoT research topics. ORCID: 0000-0001-5418-873X

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.